An Evaluation of Head, Hand and Voice Input for Menu-based Selection Tasks in Virtual Reality

(Evaluering af Hoved-, Hånd- og Stemmeinput for Menubaserede Udvælgelsesopgaver i Virtual Reality)

Jakob Worm, 202004405
Thomas Vittrup Bærentsen, 201606351

Bachelor Report (15 ECTS) in IT-product development
Advisor: Ken Pfeuffer
Department of Computer Science, Aarhus University
January 2024

AARHUS
UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

Abstract

As Virtual and Augmented Reality becomes more commercially available, a need arises for better ways to interact with the virtual world. One of the most common tasks in Virtual Reality (VR) is menu-based selection, which is why we seek to compare the efficiency and usability of various interaction modalities for completing a set of menu-based selection tasks in VR. We examine the speed with which users complete these tasks with three different interaction modalities: a head-based pointer, a finger-based pointer, and a speech recognition system activated by a head pointer. The tasks include opening information windows, purchasing items, and navigating menus. We compare the task completion speed with the users' perceived experiences across the three modalities to determine if there is a strong correlation between the efficiency of interactions and user experience potential. We find that the finger-based modality is the fastest modality of the three, and that user preference is evenly split between the finger-based modality and speech recognition, despite speech being the slowest modality. We find that the reason for this lies in the combination of the relative physical strain of the finger modality and the novelty of speech recognition. We end by proposing ways to perform a more comprehensive study in the future, with improvements to each modality and the prototype system in general.

Jakob Worm and Thomas Bærentsen,
Aarhus, January 2024.

Contents

Chapter 1

Introduction

In recent years there has been an increased commercial and cultural interest in the Reality-Virtuality (RV) Continuum [1], with Augmented Reality (AR) ranking number one in an analysis of the Gartner hype cycle for the years 2008-2017 [2]. This growing interest is exemplified by the company *Facebook*'s name change to *Meta* in October 2021, which was an effort to re-brand and declare its interest in pursuing the full commercialization and normalization of a metaverse where the physical and digital worlds are interwoven [3, 4]. Likewise, the tech juggernaut Apple has just announced that they will be entering the market with the Apple Vision Pro headset, capable of both Virtual Reality (VR) and AR [5].

Within the scientific community, numerous papers have been written on VR alone, with 1,619,054 publications as of 2023 [6]. Additionally, the scientific community holds visions of similar grandeur to the metaverse, like a future of pervasive AR as a ubiquitous and continuous context-aware experience [7]. Such a system would require a high degree of Extent of World Knowledge [1], as it would need to identify and be aware of objects and their locations in the real world. However, most of the RV Continuum is still in its infancy with relatively few people owning VR or AR headsets, and most commercial applications being primarily entertainment based, despite the potential to revolutionize how we interact with the physical and digital worlds [8].

For VR to reach its full potential we must understand the strengths and weaknesses of different interaction modalities. Our work can help identify interaction possibilities to be used across Extended Reality (ER) with a focus on how well the pure modalities work as input interactions for menu-based selection tasks within a VR environment. This knowledge is needed to choose the best interaction modalities for a given situation, which is crucial if we wish to realize the grand visions proposed by both the commercial and scientific worlds such as the metaverse or a ubiquitous context-aware experience.

We seek to quantitatively and qualitatively compare and assess three interaction modalities in a VR system. The interactions are a head-based pointer method, a finger-based pointer method, and a multimodal approach where we use speech commands in combination with a head pointer. Selection for the head and finger methods will be based on dwell time, to minimize dependent variables in the experiment. Both pointer methods

are also commonly used interaction styles within VR, so there is a need for a high level of understanding of how they compare to each other. Voice control is less used in comparison, but is still relevant within the Extended Reality (XR) scene, so we will include it in the study, as we expect it will play a larger role in XR in the future.

We include speech as a multimodal input style partly due to the famous "put that there" paper, in which a pointing-based system was combined with voice control to create a whole new range of affordances [9]. In the experiment, Richard A. Bolt manipulates a graphic interface by saying commands like "create a pink diamond there", and then providing the position using his finger. He found that the manipulation felt natural and was easy to use. We chose to use head direction instead of finger-pointing, reasoning that users naturally look toward the object they wish to interact with, making finger-pointing redundant. This leaves the hands free which in a future system could be utilized for gesticulation to improve communication [10]. The approach used in our experiment is similar to that used by other researchers. For example, Baxter et al. found that a multimodal approach using speech and gaze worked well in VR [11]. Yet both the "Put that there" and Baxter et al. studies look at speech input in isolation and do not compare it to other interaction modalities.

As for the pointing methods, we focus on pure modalities over multimodal ones, despite multimodal solutions showing advantages in effectiveness, enjoyment, and lower fatigue [12]. This choice stems from a desire to create a strong foundational understanding of the different modalities. Additionally, the scope of the paper would have grown significantly if we also looked at multimodal interactions and the permutations they would create. For this paper, we have also chosen to work solely in VR due to resource constraints, but we believe our work can be generalized and applied to some AR fields and potentially to other areas of the XR spectrum.

This thesis will consist of a related work section where relevant work is presented and discussed from the perspective of the research question. Afterward, we detail the development of an experimental prototype for testing the interaction modalities. We then go in-depth with how the different interaction modalities were created. Finally, we present the study design and the results of the study, followed by a discussion of the results and potential for future work.

Chapter 2

Related Literature

In this section of the report, we present relevant work that has shaped, reflected, or otherwise inspired the development of our thesis.

2.1 Interaction Techniques for Immersive Environments

Much has changed since the advent of AR in 1992, primarily regarding the miniaturization of computers and hardware [13]. These changes have allowed head-mounted XR devices such as the Meta Quest, Microsoft Hololens, and many more to become commercially available. Additionally, many new ways of interacting with the virtual world have become available as hardware and software have evolved.

A consensus exists on what input interactions are available in head-mounted XR, namely speech, physical controllers, hand, gaze, and head pointing [12, 14]. These all come with advantages and disadvantages that vary depending on whether the device is head-mounted or handheld [12]. Since we will look purely at head-mounted devices, we will not comment on handheld techniques or devices. Another possible input modality is a Go-Go hand-based approach, in which the user's hand extends non-linearly to interact with objects further away [15]. We will not use this approach due to the prominence of pointer-based methods in contemporary VR systems, but a comparison between it and the modalities in this thesis would be an interesting extension of this study.

Spittle et al. found that combining input interactions would lower fatigue and allow users to utilize the strength of each input while negating most of the negative aspects of said interaction, but that this comes with a steeper learning curve [12]. The paper also concluded that their findings were not definitive, as few of the analyzed papers looked at multiple interaction styles. By testing multiple interaction styles in the same study we can more accurately compare them, as different papers are bound to have different approaches to testing their interaction styles, making cross-paper comparisons harder. While we have chosen to work with pure modalities, there exists a large body of literature concerning the multi-modal input styles [11, 16, 17, 18]. In one such paper, Miniotas et al. found that a multimodal input style, which combines eye gaze and speech, has a much slower task completion speed than eye gaze as a pure modality [16].

However, this multimodal approach also resulted in a much lower mean error rate. Thus it can be argued that the choice of multimodality versus pure modality should be based on user preference or suitability for a particular task. A study on pinpointing by Kytö et al. found that eye gaze was the fastest input style, but had lower targeting accuracy than head pointing, but also that users preferred device input over gestures [19].

As evidenced by this section, we had to make choices about the interaction styles we should examine. The interaction capabilities of XR are numerous and can be combined in a multitude of ways, each with its advantages and disadvantages [12]. We chose to work with ray-casting techniques due to their prominence in the current VR landscape, as well as voice due to its ability to more abstractly solve tasks, and its potential as an input modality.

2.2 Dwell Time for Pointers

The standard for non-controller interaction in the majority of headsets is a pinch-and-click interaction using hand tracking. Many newer and luxury versions of existing headsets are, however, starting to include the hardware and software that allows for eye tracking, and thus gaze-based interactions. Using gaze-based interaction has proven less prone to error and faster than the pinch and click using a Fitts's Law task in VR [20]. A challenge pertaining to eye and head-based interactions is the so-called *Midas touch* issue, which refers to the challenge of a user activating things unintentionally when naturally looking around a scene [21]. One way to remedy this problem is to use dwell time to ensure users wish to interact with the object they gaze at, but depending on the task the appropriate dwell time is likely to vary.

Pfeuffer et al. used dwell time to add items to a cart in a virtual shopping session, and quantitative testing found that a dwell time of 1 s allowed users to complete tasks in 9.18 s compared to a dwell time of 2 s, which made task completion take on average 10.93 s [22]. The error rate with the 1 s dwell time was much higher, however, at 9.72% compared to the 2 s dwell time error rate of 2.8%. In their qualitative work, they found that users for this particular task preferred a 3 s dwell time, which interestingly also had the lowest error rate even compared to a longer dwell time of 4 s. For our prototype, we chose a dwell time of 1 s for purchasing tasks, since the error rate is not an important factor in our test. The dwell time for interaction is otherwise 0.3 s since it was found appropriate in testing.

2.3 Alternative Input Interactions

Eye gaze for selection is gaining purchase within XR. There are plenty of examples of eye gaze being used for selection tasks or to support selection [17, 23, 24]. Work by Blattgerster et al., found that eye gaze outperforms head pointing selection in speed, task load, head movement, and user preference [25]. In contrast, a study by Qian et al. found the head selection to be faster and less prone to error than eye interaction using a Fitts's Law test [26]. Additionally, they found that users had a strong preference for head-based selection [26]. Since we do not have the option of implementing eye tracking, we will

only examine the previously mentioned head, finger, and voice modalities. However, should the scope of the study be expanded, gaze selection would be a natural extension we will discuss this in Section 7.3.

Recently there have also been forays into using multimodal interactions to enhance usability and user enjoyment. One such example is the work by Lystbæk et al. on gaze-assisted selection-based text entry, in which they created a prototype where users would test four approaches to text entry. Two of the prototypes presented in the paper were multimodal, using both gaze and hand. They compared these with the pure modality approaches of air tap and dwell time and found that users preferred gaze-finger interactions, although the study concludes that each of the presented methods had both pros and cons [18]. In another paper, Lystbæk et al. combined eye Gaze and mid-air pointing for interaction with menus in augmented reality and found that gaze-assisted techniques outperformed hands-only input [17]. Another conclusion drawn in the paper is that eye gaze works almost like the hover state known from regular 2D GUI interactions.

As mentioned previously, we are unable to use gaze-based interactions due to resource constraints, and will instead use head pointing both as its own modality and in combination with speech recognition. We will focus on studying pure modalities rather than mixed ones, to provide a strong base for comparison with more complex solutions.

2.4 Adaptive Interfaces

One way of designing an interface in VR is to use an adaptive approach where the content reacts to the user. A study by Piening et al. indicates that people preferred gaze adaptive interfaces over an all-on approach to UI design, and further suggests some design rules for gaze adaptive AR [27]. The findings show the importance of indicating what objects can be interacted with, for instance with a graphical marker. Furthermore, it is suggested that these markers react when interacted with and that the reaction occurs quickly, within 0.2 s to 0.5 s. Following these design principles, we utilize markers for interactable elements in the prototype. These markers are small eye icons seen in Figure 2.1 regardless of interaction. Additionally, we use a slider to indicate the reaction on the markers. Once the slider fills the menu will open thus adapting to the user, as can be seen in Figure 3.3.

Figure 2.1: In the prototype, a small eye symbol indicates that the object can be interacted with.

Another approach can be seen in AR where it is possible to anchor UI elements to either a real-world object, some part of the user's body, or some combination of the two. This has been shown to elicit positive responses from users in a study by Lu et al., who found that the form factor of the head-mounted devices was the biggest barrier to everyday use [28]. We took some inspiration from this AR approach and have anchored menus to their virtual counterparts.

A different paradigm within AR is that of glanceable AR, in which information stays at the periphery of the user's vision, and is thus accessible with a glance whenever needed [29]. This is in contrast to other solutions where the information is presented as if it were a part of the real world. This approach is less obtrusive than other options in everyday use cases [28]. In a study by Davari et al., they sought to validate the need for glanceable AR [30]. To this end, they created two scenarios: one solo, in which the user was doing some real-world activity alone in a static environment, and one social, designed to represent an authentic social conversation. In both scenarios, the glanceable interface proved more effective at information gathering and ranked higher in user experience than using a smartphone [30]. They argue that glanceable AR apps have the potential to replace smartphones as the primary everyday information access tool. As the name implies, glanceable AR requires eye tracking and can therefore not be implemented in our prototype.

We have utilized some of the techniques commonly used within adaptive interfaces for our prototype, for example in the use of markers as seen in Figure 2.1. Additionally, we resize and change the orientation of information boxes depending on the distance and viewing direction of the user. We have also taken great care in the placement of UI elements, making sure they do not obscure the virtual objects they belong to. We have chosen to design our interface so it can easily be adapted to an AR setting because we in Section 7.1 want to discuss how or if our findings can be applied to adaptive interfaces within AR.

2.5 Label Placement in Augmented Reality

A significant challenge in AR is known as *the label problem* [31]. The problem concerns where to place labels (and by extension other elements) in an AR environment without creating overlap between different labels, or between labels and real objects. To this end, researchers have defined metrics to be used for label layouts to maintain visual clarity and coherence between the co-referential elements of the textual and visual elements [32].

When determining label placements in VR and AR, a strategy of continuous view management can lead to poor temporal coherence since the viewpoint constantly changes with the user's head [33]. This can be alleviated by using discrete view management instead, where the label placements are static and only change when the viewpoint moves significantly from the initial point, thus trading potentially inferior layouts for better temporal coherence. This approach is supported by the work of Lindlbauer et al., who found that the only way to handle the constantly changing context of a dynamic system is an automatic adaption of the information presented to the user [34]. To support the development of adaptive User Interfaces (UI), Belo et al. created a system adaptive user interface toolkit (AUIT) for XR [35]. The toolkit allows developers to easily build adaptive interfaces without being weighed down by repetitive and time-consuming programming tasks, by allowing them to specify adaption objectives to which the UI will conform. To deal with the label problem we have chosen to have static labels centered on interactive objects. Since users are placed in the middle of the scene, all labels are visible from their vantage point.

2.6 AR in the Tourist Sector

Literature shows that introducing AR to tourist destinations makes it possible to increase their competitive advantage compared to other destinations [36]. This is exemplified in a study by Alzua et al. where they found that tourists felt the experience of using stationary binoculars to look at a town in northern Spain was enhanced by using AR [37]. When looking through the binoculars the tourists could see markers around certain cultural sights around the city. They could then open a menu with options using buttons on the side of the binoculars.

Another approach to AR-based enrichment of city spaces can be seen in Ingram's work where he proposes a design for a guide to work in "augmented cities" where users can annotate objects and where annotations are moderated using trust-based filtering [38]. Based on these papers we have chosen that our prototype should be set in a tourist setting as this would be a realistic real-world application. The Virtual Environment has several "landmarks" in the form of the equestrian and Discobolus statues that are intended to create an experience akin to being in a tourist destination, as well as other interactable elements which can be seen in Figure 3.1.

Chapter 3

Prototype Design and Concept

In this section, we describe the concept and development of the prototype used in this study. The purpose of the prototype was to create a virtual environment wherein different interaction modalities could be tested on the completion of menu-based selection tasks. We have chosen to set our prototype in a city that could work as a tourist destination, based on the literature described in Section 2.6. We designed the scene so that the user would be standing in the middle of a road surrounded by various landmarks, which can be seen in Figure 3.1.

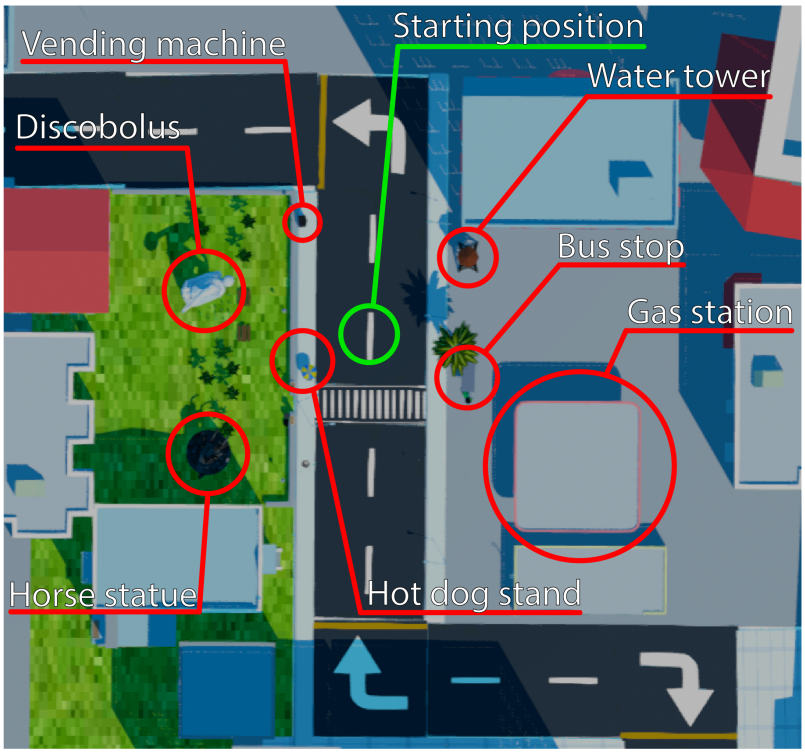


Figure 3.1: A bird's eye view of the interactable elements in the scene, where the user starts in the center.

3.1 The Prototype Construction

The prototype is built using Unity version 2021.3.17f1 [39] for the Meta Quest 2 VR headset using the Oculus Integration for Unity package version 3.2.3 [40].

We built a small virtual city using assets from the Unity Asset Store to simulate a tourist destination [41]. This gives us complete control of the objects in the city, allowing us to mark the interactable objects.

The user starts in the middle of the street to have the best view of the environment on either side. To the left, there is a park surrounded by a low stone wall, and in the park, there are two large statues. One is of a rearing horse [42], and the other is Discobolus [43, 44]. Outside the park, there is a hot dog stand and a vending machine. On the other side of the street, there is a water tower, a bus stop, and a gas station with two cars. A top-down view of the scene can be seen in Figure 3.1.

(a) View of the park with the two statues.

(b) View of the bus stop and gas station.

(c) View of the water tower and the task text.

Figure 3.2: Views of the interactable elements in the city environment.

3.1.1 Interactable Elements

The scene features eight interactable elements that can be seen in the top-down view in Figure 3.1 and from the user's perspective in Figure 3.2. The elements all have information boxes that open in the same way, but the contents of the boxes vary for each element:

- The **vending machine** has a list of available items, but they are not purchasable in the study.
- The **Discobolus and horse statues** have boxes with general information and a slideshow of pictures the user can navigate.
- The **hot dog stand** has a menu where the user can buy different kinds of hot dogs, as well as two sub-boxes. One has some reviews (which are not used in the study), and the other has pictures of wares the user can navigate.
- The **water tower** has a box with general information and no interactable features.
- The **bus stop** has a list of tickets for different buses which the user can purchase.

It also has a sub-box with a countdown to the arrival of the next bus (the bus will not arrive in the scene).

- The **gas station** has two cars with information boxes and no interactable features.

3.1.2 Interactions in the Scene

Opening Information Boxes

The most common interaction in the scene is the opening of information boxes. These can range from information boxes containing relevant information about the viewed object to more interactive ones featuring the ability to purchase an item or view and navigate pictures. When using either of the two pointer methods, information boxes are opened by focusing the ray at an object with the eye marker seen in Figure 2.1. A circular slider will then appear to indicate that the object is being focused on, and gradually fill. Once the slider is full the box opens. The information boxes close automatically after 3 s of not being looked at or pointed at. Boxes that are considered sub-boxes such as picture boxes and review boxes will be opaque until focused on by the pointers, to reduce the amount of information presented at once. To open a box with the voice, you have to look at an interactable object and utter a phrase containing the word "open", or something similar. To close the box you must similarly tell the object to "close".

(a) The slider is partly activated, and there are no information boxes.

(b) The slider is fully activated, and the information boxes have appeared.

Figure 3.3: The process of opening an information box using the slider.

Viewing and Navigating Picture Boxes

Some information boxes have a picture sub-box connected to them. This box will display item-appropriate pictures and allows the user to navigate through the picture gallery using either the arrows at the bottom of the box if using pointers, or commands like "next picture" if using voice. This limits voice to some extent - as described in

Figure 3.4: The sub box with pictures is visible, while the review box below it is still opaque.

Section 4 - as it forces the voice modality to navigate pictures similarly to pointers instead of using more abstract phrases like "show me the picture of a man eating a hotdog".

Purchasable Content

Some boxes in the scene, namely the hot dog stand and bus stop, have the option to purchase either a hot dog or a bus ticket. With pointers, this is done by hovering the purchase button over the appropriate item and then hovering over the yes button to confirm the purchase. The buttons will turn blue when hovered to provide visual feedback. They will also play a sound for auditory feedback. If the user hovers over the no button then the popup simply closes again. If an item is purchased a popup will appear in front of the user telling them which item was bought. This process can be seen in Figure 3.5. With voice commands, users can bypass the buttons on the boxes and simply ask to purchase an item as long as the appropriate box is open, thus allowing the voice modality to work more efficiently.

(a) Choosing to buy a bus ticket.

(b) The box asking to confirm the purchase.

(c) The popup indicating the ticket was purchased.

Figure 3.5: The process of purchasing a bus ticket.

3.1.3 Sound

The scene features simple audio cues when interactable features within the information boxes are hovered. These sounds serve to give a sense of polish to the system since it does not feature other sounds, and since many programs utilize sound for emphasis we expect the study participants will expect it to some extent.

3.1.4 Information Box Resizing

The sizes of the information boxes are set manually, but they will adjust their size in the scene relative to the position of the user to always have a visual angle of 60°. With the prototype, this is a bit redundant since the study participants don't move around, so we could have sized and orientated all canvases toward the user. But with the visual angle, we allow the system to function even if the scope of the study is expanded to include an area for participants to move around in.

3.2 The Input Modalities

This section describes the functionality and implementation of the three modalities in greater detail.

3.2.1 The Head Pointer

In the head modality, two invisible rays are drawn from the center eye anchor of the headset, which represents a point slightly above the midpoint between the user's eyes. These rays let the user open the information boxes of interactable elements by dwelling on them for 0.3 s, as well as interacting with the information boxes in the same manner. One ray handles the activation of interactable elements while the other handles the interactions in the information boxes. Both rays are overlapped so that they will look and feel like a single ray to the user. To close an information box, the user must simply look away from it for 3 s and it will close by itself.

3.2.2 The Finger Pointer

The finger modality works in the same way as the head modality, except the rays are drawn in a line from the first knuckle of the index finger to the tip of the index finger. In this modality, the rays are visible as a simple green line to provide better visual feedback.

3.2.3 Head Pointer and Voice Control Input

We have implemented speech commands using the natural language Artificial Intelligence *Wit.ai*, which integrates with the Voice SDK and the Oculus Integration SDK [45]. Wit.ai is a natural language model developed by Meta, which can be trained on the website by writing "*Utterances*" and then coupling them with an "*Intent*" [46]. The website then allows for training of the AI and lets you validate the outputs. After training you can activate Wit in your Unity app and it will record what is said and send

it to the Wit.ai API. The server then finds the intent from the recorded utterance and, if it is within the scope, sends the intent and potential entity values such as numbers or names back to the app. Specific intents are coupled with functions that the app will execute upon receiving an intent.

In our prototype, the speech recognition is activated by the use of a head pointer, and the system remembers the last interactable element the user looked at, which it uses as a reference for commands such as "open" or "close". This use of a head-pointing activation means that the speech modality is not entirely a pure modality, unlike the finger and head modalities, but Wit.ai allows it to be close to a natural language system in many aspects regardless.

Chapter 4

User Study of VR Interactions in a City Environment

Our study aims to investigate the speed with which participants can complete a series of simple tasks using the three different modalities, as well as their satisfaction with each one. The interaction styles will be compared in terms of task completion speed and their score in a NASA-TLX questionnaire [47, 48]. We took great care in choosing the tasks as certain interaction styles are more suited to particular types of tasks [12]. Since we seek to test the interaction styles in comparison to each other, we must be aware of task biases. For this, we looked at Spittle et al. who categorize appropriate inputs to distinct tasks. They found that speech worked well for abstract tasks, while head and finger pointers worked better in menu selection tasks. Our tasks are split into three types which are based on menu selection tasks. We are aware that this may give an advantage to the pointer-based modalities, but the tasks are fitting for a realistic system, which the voice modality must also fit into.

- **Information seeking:** The participant must find and open the information box of a specific target in the scene. The participant is not asked to locate any information in the box, as this would introduce a variable of reading speed. The pointer interactions open the menus based on dwell times while the voice modality works by an "open" command. The latency of the voice commands exceeds the dwell timer, which gives the pointers a slight speed advantage.
- **Picture seeking:** The participant must find a specific picture attached to a relevant target in the scene. These tasks are designed to cause the participant to interact with some simple controls in the UI. Again the pointers have a short dwell time between transitions which outperforms the latency of voice controls, and since the task often requires going through multiple pictures, we believe that the pointers will have significant advantages in these tasks.
- **Purchasing:** The participant must purchase an item from a specific target in the scene. This is also intended to make the user interact with the UI controls. These tasks require the pointers to navigate sub-menus whereas voice is allowed

to bypass them by using more abstract commands, thus giving it an advantage.

The participants will each complete three tasks of each type for each of the three modalities, for a total of 27 tasks. Since we tested the prototype with 12 participants this gives us a total of 324 tasks performed. Whenever a participant feels they have completed a task, they will give a thumbs-up gesture to indicate that they are finished with the task. This method of recording completion introduces a flat delay to the time of every task, but since the study aims to compare relative times between the modalities, this should not impact the results. Furthermore, using a self-reporting method ensures that the task will only be marked as completed once the participant feels it is. Since the study also includes the participants' experience with the modalities, they mustn't risk accidentally completing tasks without realizing it, or feeling that they should have completed a task while the system disagrees.

We will record the time taken to complete each task to be used for quantitative analysis. In addition, after completing each modality the participant will fill out a NASA-TLX questionnaire about their experience with that modality.

The research questions for the study are as follows:

- **User performance:** How quickly do the participants complete each kind of task with the different modalities? This may vary between users, but we hope to see one or two of the modalities be noticeably faster than the others.
- **User experience:** Which modality has the highest user satisfaction? Even if a specific modality is significantly faster than the others, the participants may not experience it as such or may find it straining to use. Since this study aims to illuminate which type of interactions would be ideal for an applied VR or AR system, the user experience may end up being even more important than quantitative performance.
- **Preferred interaction modality:** Which modality do the users prefer in their own words? Aside from the TLX results, the participants may have a favorite modality for reasons not apparent from the rest of the study. Responses to this question could shed light on otherwise unexamined factors of the system.

4.1 Experimental Design

The tests were completed one at a time, with the participant being informed of the structure of the test before starting. The introduction included information about the three kinds of tasks, the method for indicating task completion, and a description of the first modality they would be using. While the participants were performing the tests, we were available to answer any questions they might have regarding the interaction possibilities or the tasks themselves.

To eliminate order effects for the modalities we used a Latin square of the three modalities to determine in which order the participants would complete the modalities [49, 50]. To eliminate order effects for the tasks we instead used a random order, since the larger number of tasks made using a Latin square infeasible, while making the

randomness over all 36 tests more likely to have a flat distribution. Each time the program runs, it picks three of each kind of task and then shuffles the list, ensuring a random distribution.

4.2 Statistical Analysis

We use SPSS Statistics for our qualitative data analysis [51]. For task completion speeds we first make a Within-Subjects Effects table to determine if there is an overall statistically significant difference between the mean completion time of the different modalities.

We then use a Mauchly Test of Sphericity [52, p.1] to see if our data violate the assumption of sphericity to determine if we have to use a Greenhouse-Geisser correction [52, p.2]. If the difference in task completion speed between modalities is found to be statistically significant ($p < 0.05$), we use a Bonferroni post hoc test to discover which specific modalities have a statistically significant difference in task completion [53].

For the NASA-TLX assessment we perform a Friedman test in SPSS Statistics, which is used to test for differences between groups when the dependent variable being measured is ordinal [54]. We perform the Friedman test for each question in the questionnaire to discover if there is a statistically significant difference in how users rate that attribute across the modalities. If a significant difference is found, we use a Wilcoxon signed-rank test to see for which modality pairs the difference occurs [55].

4.3 Participants

We recruited 12 volunteers for the user study (6 male and 6 female). Participants were between 19 and 29 years, and 9 out of 12 were IT Product Development students from the Aarhus University Department of Computer Science. Out of the recruited participants, 4 were familiar with VR. The study took place in a classroom at the university with one participant arriving at a time and under the supervision of one experimenter, who introduced the premise of the study. We will discuss the limitations in regards to participants in Section 6.4.

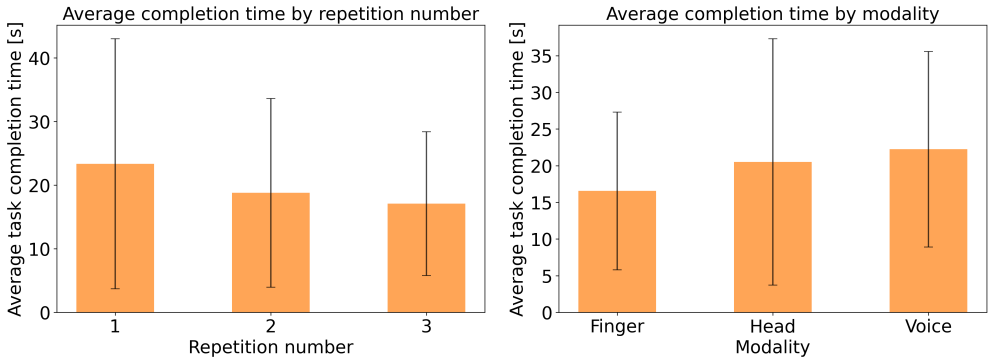
Chapter 5

Study Results

In this section, we present the result of our user test. First, we present the task completion speed results, followed by the NASA-TLX questionnaire results, and lastly the qualitative results of the user test.

5.1 Task Completion Speed

The average time to complete a task based on the repetition number, shown in Figure 5.1a, clearly shows how the participants need some time to adjust to the system before they use it effectively, regardless of modality. This is counterbalanced by using a Latin square to determine modality order, as discussed in Section 4.1.



(a) Completion time by repetition.

(b) Completion time by modality.

Figure 5.1: The average task completion times ordered by repetition and modality, with standard deviation error bars.

As seen on the histogram in Figure 5.2, there are some significant outliers in the task completion times. This is consistent with our observations during the tests, as the system would sometimes double-register a thumbs-up gesture, and instantly complete a task.

To account for this, we remove outliers below two seconds, accounting for a total of 4.42% of the data.

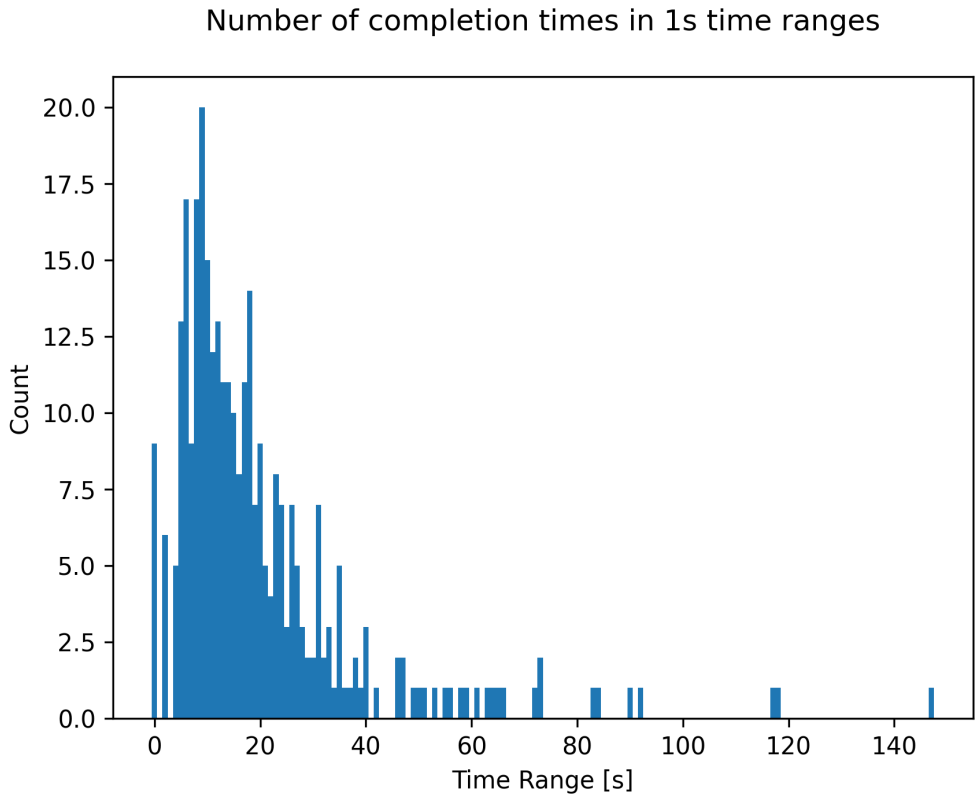


Figure 5.2: Histogram showing the number of tasks completed in 1-second intervals.

The average task completion speed for each modality can be seen in Figure 5.1b, along with standard deviation error bars. The figure shows that the finger modality was fastest on average, followed by the head, and finally voice. It should be noted that the head modality had a larger standard deviation than the others.

For a statistical analysis of the task completion time we created a Within-Subjects Effects table using SPSS, and with it we were able to find the f value for the time factor, its associated significance level, and effect size (Partial Eta Squared). We then checked if our data violated the assumption of sphericity using Mauchly's test ($\chi^2(2) = 0.678, p = 0.713$), and found that it did not. Then we used an ANOVA with repeated measures over the mean scores for the modalities' task completion times and found that they were statistically significantly different ($F(2,20) = 6.724, p < 0.006$). To see between which modalities the statistical significance occurred we did a post hoc analysis with a Bonferroni adjustment [53] which revealed that finger task completion was significantly faster than voice task completion ($p < 0.014$).

5.2 NASA-TLX

As described in Section 4.2, we examined if there was a statistically significant difference between how the attributes were experienced by the participants across modalities. The average values of the responses are shown in Figure 5.3.

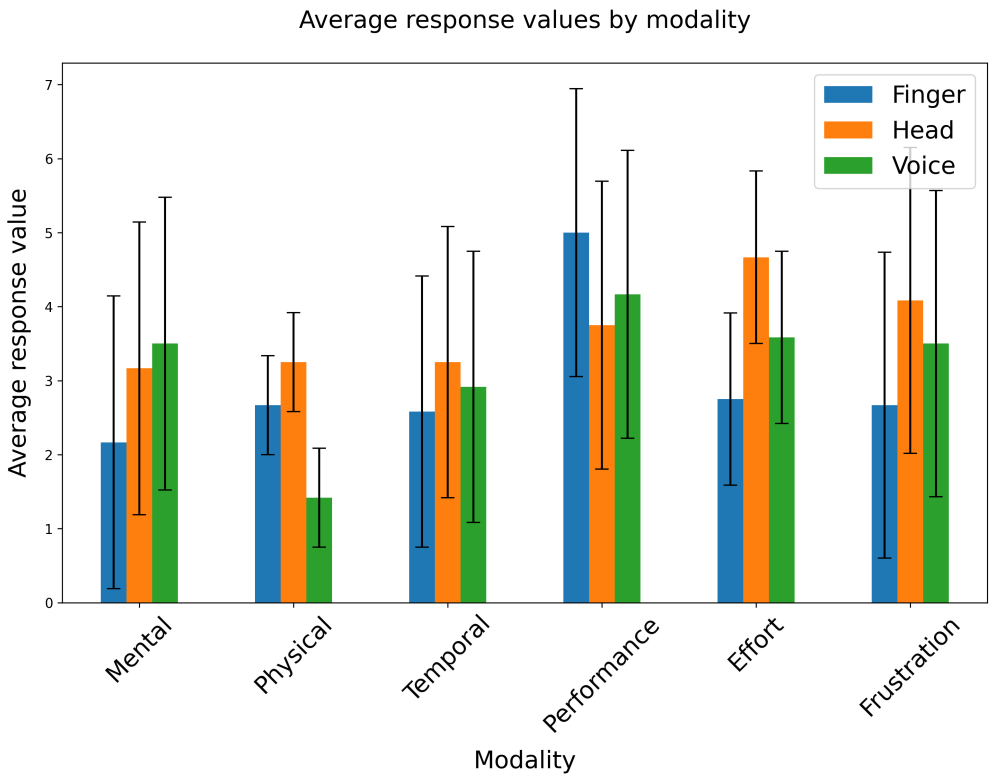


Figure 5.3: The average values of the NASA-TLX responses by modality, with standard deviation error bars.

There was no statistically significant difference in mental demand depending on the modality ($\chi^2(2) = 5.200, p = 0.074$). The median responses were Finger = 2.00, Head = 3.00, and Voice = 3.00. There was a statistically significant difference in physical demand depending on the modality ($\chi^2(2) = 11.744, p = 0.003$). A Wilcoxon signed-rank test showed that there was a statistically significant difference between the voice and finger modalities ($Z = -2.588, p = 0.010$), and between the voice and head modalities ($Z = -2.701, p = 0.007$). The median response was 1.00 for voice and 3.00 and 3.50 for finger and head respectively. There was no statistically significant difference in temporal demand depending on the modality ($\chi^2(2) = 0.812, p = 0.666$). The median responses were Finger = 2.00, Head = 3.00, and Voice = 2.00. There was no statistically significant difference in perceived performance depending on the modality ($\chi^2(2) = 2.229, p = 0.328$). The median responses were Finger = 6.00, Head = 3.50, and Voice = 3.50. There was a statistically significant difference in perceived effort depending on the modality ($\chi^2(2) = 12.762, p = 0.002$). A Wilcoxon signed-rank test showed that there was a statistically significant difference between the head and

finger modalities ($Z = -2.589, p = 0.010$), and between the voice and head modalities ($Z = -2.511, p = 0.012$). The median response was 6.00 for head pointing, and 4.00 and 4.75 for the finger and voice respectively. There was no statistically significant difference in frustration depending on the modality ($\chi^2(2) = 4.333, p = 0.115$). The median responses were Finger = 2.50, Head = 4.00, and Voice = 3.00.

5.3 Qualitative User Study

In addition to the standard NASA-TLX questionnaire, we had participants answer two questions for each modality:

> *What did you like about this modality?*
>
> *What did you dislike about this modality?*

And a final question after all modalities and questions were completed:

> *What is your favorite interaction modality and why is this your favorite?*

As the participants are anonymous we will refer to them as P1-P12. We found that most preferred either voice or finger pointing, with only two participants choosing other options, as seen in Figure 5.4.

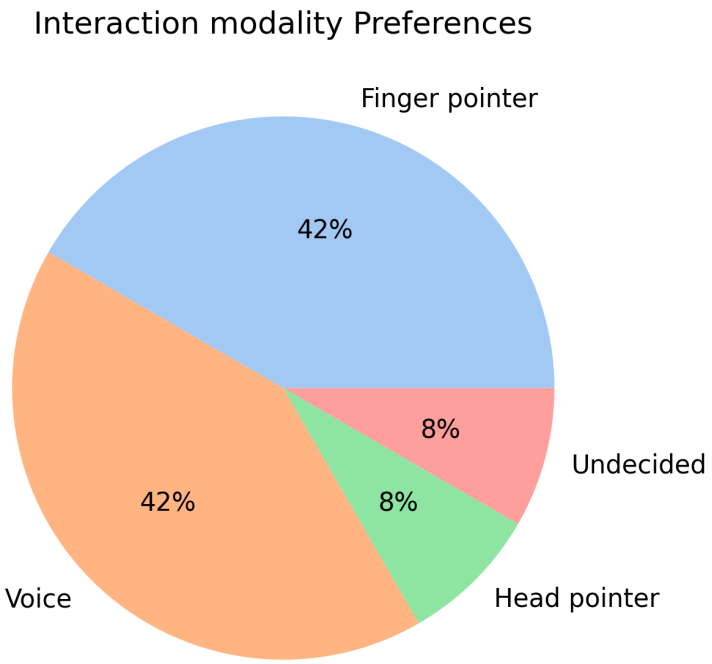


Figure 5.4: The distribution of the participants' favorite interactions.

When asked why they preferred head interaction, P10 answered: "*Mostly because other people wouldn't know what I was doing, or be able to see it from the outside.*"

We hypothesize this sentiment would be more common if the experiment had been conducted in a more public setting such as the wild. Another user, P7, stated they did not like to use their voice possibly due to the same privacy concerns as P10: "*Hand interaction was the most intuitive so I like it the most, though it also had issues. I also don't want to talk to interact so that at least rules out the speech modality.*" As we only saw the results afterwards we could not verify if P7 did not want to talk due to privacy concerns.

When asked what the participant liked about the head pointing, P8 responded: "*It makes sense that when I look at a thing the information pops up. There is no doubt about where you can find information if it automatically pops up when you look at it.*" As for what participants disliked about the head, P6 said: "*When you enter the menus you have no idea where you are. it's very difficult to select the items you want. When finding the pictures I felt like it was more a matter of luck than me actually looking through them. The same when buying bus tickets.*" Both of these sentiments were shared among the other participants.

Participant P6 preferred voice, saying: "*It was much faster and easier to navigate the smaller menus. It also made sense to me since pointing for me feels a little like trying to touch something which is contradictive in VR.*" Another participant, P9, who also preferred voice said: "*I liked the voice interaction the most. It was most easy for me and less physically demanding. I was not doing wrong as many times, as the others.*" Here we see how a user noticed and commented on the relatively lower physical demand for voice interactions. This is seen again with participant P5 who said: "*I liked the finger-pointing interaction the most, it felt natural and easy to complete my tasks. The strain was not too much, however, long-term use might leave you tired.*"

When asked about what they liked about voice commands, P4 said: "*it was good at registering my responses despite me not speaking very loudly or clearly normally.*" and P11 said: "*It was a natural interaction, so it was fairly easy to have the system do what you wanted. It was surprisingly efficient at hearing what I asked it to do.*" When asked what they dislike about the voice modality, P2 said: "*Sometimes the images changed slowly, so I was not sure if the system had recognized my voice*". P8 said: "*If you are not familiar with voice-activated interfaces (Siri), you can get a little insecure. I think a small tutorial where it ask you to say specific commands to show which words work.*" Another participant, P11, was also unsure about what the system responded to, saying: "*I did not know which prompts the system would accept so I felt silly asking it to perform a task when I did not know if it understood what I said. I forgot to close the dialogue boxes because it seemed sort of natural to me that that would happen automatically.*"

When asked what they liked about Finger pointing, P1 said: "*It was very easy and quick to do. There was an indicator for where you are pointing*" As for dislikes P6 said "*It's difficult to be precise with it. If your hand is shaking it makes it difficult to hit small targets. Especially when you have to hover over it for a certain time to choose it.*" and P7 said: "*It was difficult to be accurate at long ranges(bus tickets...)*".

P11, who preferred finger pointing, answered: "*I liked the [finger] pointing interaction the most because it felt natural to me given the circumstances I was in. Even though it*

felt more "natural" to speak and use gaze for selection, it was also more difficult because that is not how I usually interact with technology.". We did not expect an unfamiliarity with speech technology due to its prevalence in phones and home appliances like google home, but this indicates an additional potential challenge for the voice modality.

Chapter 6

Discussion

In this section, we discuss the results of the user study and the issues that appeared during the development and testing processes.

6.1 Comparison of Modalities

Looking at the average task completion speed we see that voice ranks lowest among the modalities (Figure 5.1b). Despite this, the qualitative comments presented in Section 5.3 show that this is one of the interactions people liked the most, possibly due to it being less physically demanding than the head and finger modalities and requiring less effort than the head modality, as seen in Section 5.2. Looking at the qualitative responses we see that some participants found it very natural to use the voice while others were more unfamiliar with voice control and natural speech. We also noted a high number of responses commenting on the fidelity of recognition. As expected and discussed in Section 4, the picture navigation task was difficult to complete using the voice modality as seen by the answer given by P2, who began to doubt if they were doing it right.

We see the finger pointer had the highest physical demand, but despite this it was still one of the most liked modalities, likely due to it ranking the lowest in effort, which can also be seen in its fast task completion speed. The high physical demand for finger-based pointing is in line with the findings of Spittle et al. [12]. As seen in Figure 5.1b and Section 5.1 this modality was faster than the other two, and statistically significantly faster than the voice modality. We also saw from the qualitative work that participants had trouble when interacting with objects further away from the user despite the visual feedback in the form of the line.

While it was the least preferred among participants - with only one participant citing it as their favorite - the head-pointing modality was not statistically significantly faster or slower than the other modalities. It did, however, rank the highest on physical demand, temporal demand, effort, and frustration. Thus it is perhaps not surprising that only one participant, P10, preferred the head modality. This is further backed up by P10 citing

privacy as the reason for their preference, which is an aspect that is not represented in the TLX.

Based on these results, our impression is that the finger modality is generally better than the others at menu-based selection tasks, based on its superior task completion speed combined with its low effort rating. As seen in Figure 5.4, it is also one of the most popular modalities which will be discussed further in the next section. The question is whether users value low physical effort higher than speed, in which case the voice modality would be better.

6.2 Overall User Preference

As seen in Figure 5.4, the users were generally split between preferring the finger pointer and the voice modalities. There may be some degree of bias to the results, as the head modality was generally regarded as the least polished, based on our opinions and informal user feedback during and after the tests. Specifically, the pointer was invisible and originated slightly above the eyes because of the way the app tracks head position. This made it difficult to consistently hit smaller targets such as the buy buttons, which caused frustration.

Another point that may have biased the results is that many users seemed impressed by the speech recognition capabilities of the voice modality, potentially causing them to enjoy that modality more over the short span of the test. Besides potentially influencing the TLX responses, this result could be interesting by itself. Since VR and AR are still fairly new technologies a lot of people are not used to being in Virtual environments, and since voice control is still a relatively new and unfamiliar way of interacting with computers for some, this experience might have been rated as more novel than the other modalities. With the recent popularisation of natural language processing by ChatGPT [56], we believe a lot of people will find natural speech to be a more familiar way of interacting with computers in the coming years, which means that some of the novelty of natural-speech interaction with a VR or AR system may disappear, but be replaced by a familiarity that may ultimately lead to better user experience than more traditional solutions.

6.3 Comparison of Qualitative and Quantitative Results

As described in Section 6.1, the measured task completion speed of the modalities did not directly correspond to the perceived user experience. For example, the finger modality was the fastest of the three but was also reported to have a higher physical strain than the voice modality. This seems to support the findings of Spittle et al. who also cited the physical strain of finger interactions [12]. To gain more insight into this, we would need to perform a second study in a real location and for a longer duration. We will discuss this possibility further in Section 7.2.

The results imply that task completion speed may not be the deciding factor when choosing which interaction modality to implement in an AR or VR system. This is supported by the work by Kytö et al., who found that users preferred device input despite

eye gaze being faster [19]. This means that developers may have to choose whether to prioritize interaction speed or user experience, since the perceived effectiveness of the system may not correspond to the more objective measures.

6.4 Limitations of the User Study

The user study included only 12 participants, which is few enough that any outliers can significantly affect the outcome. We also made no effort to include participants with different backgrounds or levels of familiarity with technology in general or VR in particular. As such, our findings may change if the study is repeated with a group more representative of the general populace. A study by Faulkner showed that a usability test with 5 participants found 99% of problems, while other groups of 5 only found 55%. The worst performing group of 20 users found 95 percent of problems, however, [57]. This means that it is likely we would find more flaws in the system with a larger user base.

We also noted in Section 6.2 that the novelty of the speech recognition may have biased the TLX answers to some extent. This could be averted by performing the study over a longer time, to allow participants to become familiar with the technology. A longer-running study could also exacerbate issues such as the physical demand, which, according to Section 5.2 would make the voice modality preferable. Likewise, performing the study in a real location (presuming the system is converted to AR) could impact parameters such as effort, frustration, or temporal demand, since the participants would have to keep in mind both the real world moving around them and their current tasks. The fact that one user reported head as their preferred modality because of its privacy aspect could also indicate that more users would dislike having to speak aloud or make large arm movements in a public space.

The prototype itself also had some significant limitations caused by time constraints in the development process. As mentioned previously, the head pointer originated from a point above the user's eyes and was invisible, which led to some difficulty in hitting smaller targets. The speech recognition was limited by having to wait for a response from the server, and by having to activate the listening instead of having continuous listening. These factors combined added a noticeable delay for the voice modality, which might have made it more cumbersome to use.

The tasks themselves could also be expanded to explore the modalities in greater depth. The tasks which included opening an information box were kept simple because of the aforementioned delay in speech recognition which made larger boxes very difficult to navigate. A larger study with an improved prototype should thus include larger and more varied tasks, perhaps with multiple stages to each task to better simulate a realistic use case.

6.5 Voice Control as a Mixed Modality

As stated earlier in the thesis the initial idea was to test and compare the pure modalities by their own merits. The voice modality, however, uses head pointing in combination

with speech due to technical limitations with the interaction. This led to some confusion among the participants when they tried to open a specific panel, and a different one would appear because they had not looked in the right direction.

There was also some confusion regarding what kinds of commands the system could recognize. Some participants attempted to simply ask for the picture they were tasked to find, only to be disappointed to learn that the system only recognized requests for the next or previous pictures.

Overall the participants still liked the voice modality, but these missteps have likely soured the opinions of some participants. In Section 7.3 we will discuss ways to turn the voice interactions into a pure modality to remove error sources and to more closely reflect on the purpose of the study.

Chapter 7

Future Work

In this section, we suggest ways in which the study could be improved and expanded upon in the future. The suggestions are based on the results of the study and the discussion topics presented in Section 6.

7.1 Moving to Augmented Reality

To say with certainty that the findings from this paper apply to AR as well as VR, it would be necessary to recreate the study with an otherwise identical AR system. With the necessary hardware, this should not present a major challenge, given the design of the system which depends on invisible interactive areas in the world space. For the move to AR, it would only be necessary to specify a position for the participant to stay in, and then place the interactive areas over real-world objects.

More likely, however, the switch to AR would be combined with some or all of the changes to the study discussed in Section 7.2, in which the participants will move around in a more realistic scenario. Since this would allow multiple different perspectives of the interactive elements, it would necessitate the implementation of a computer vision system to identify the interactive elements and move their information boxes accordingly, as well as to fetch information about the elements, giving the system a high Extent of World Knowledge [1]. This would also require the implementation of some of the techniques described in Section 2.5, to ensure a cohesive layout of the information boxes.

7.2 Improved User Study

In a second, improved study, the issues described in Section 6.4 could be addressed. Besides improving the fidelity and performance of the prototype, such as by fixing the issues with the head pointer mentioned in Section 6.2, recruiting a larger number of participants would allow for much greater certainty from the statistical analysis. More participants would also allow us to reach a broader spectrum of participant backgrounds, which would eliminate bias from factors such as preexisting familiarity with VR.

Developing to prototype into an AR system as described in Section 7.1 would also allow the study to be moved into the real world, including the challenges that would bring for the participants. Requiring that the participants stay conscious of their surroundings while performing the tasks could greatly influence the results, as could the self-consciousness that comes from being around strangers while performing the tests. Such a study could also involve tasks that require the user to physically move around an area, such as buying a bus ticket and then taking the bus.

Finally, a longer-running study could make previously unnoticed factors apparent. Some modalities may become easier to use with practice to a greater extent than others, while the physical strain of some could become more pronounced over longer periods. The longer-running study could include longer sessions of tests, which would highlight the physical effects, and a multi-day study, which would highlight the effects of greater familiarity with the system.

7.3 Improvements to the Prototype

With the switch to AR hardware described in Section 7.1, it would become possible to implement gaze as an interaction modality, since most AR-capable headsets support eye tracking. This could be done as a replacement for head tracking, which was almost unanimously the least liked, or as a complementary fourth modality. A reason to include both gaze and head in the same study would be to either refute or affirm the results by Qian et al. who found head pointing to perform better than gaze in task completion speed and user preference [26]. The results from including a gaze modality would also serve as a point of comparison to findings by Lu et al., who evaluated two gaze-based systems, and found that users appreciated a glanceable design [28]. Since eye tracking is typically much less exact than simple raycasting, it would likely be necessary to modify the layout of the information boxes, for example moving the buttons further apart. Alternatively, eye tracking could make its inputs based on the most-watched object over a time frame, which negates the effects of the relative inaccuracy of the eyes [19]. Either of these solutions would likely be inferior to the adaptive interfaces used by Lu et al., but to make a baseline comparison between the modalities, the interfaces need to be identical for all the modalities.

As described in Section 6.5, the voice modality was somewhat limited in scope compared to its potential. An improved version of the prototype could do away with the head pointer for this modality and instead implement a more comprehensive natural speech system, which would allow users to interact with objects they are not currently looking at. This would allow the participants to complete several of the tasks significantly faster, since they would only need to state their intent to, for example, buy a bus ticket, and could skip the intermediate steps. Because of this, it would be necessary to reevaluate the tasks, and possibly include some which would be slower for the natural speech modality, to achieve an unbiased study.

A more advanced prototype could include multimodal approaches and voice as a single modality for potentially deeper insights into which specific parts of each modality the participants respond well towards. Multiple papers have shown the potential benefits of

multimodal approaches, so we could likely find one or several which worked better than the single modalities in this paper [12, 16, 18]. Furthermore, a multimodal prototype could be used to compare results between approaches that feature the same modalities, and thus find specific elements of modalities which users particularly like, or use cases in which they are especially fitting. With this approach, it would also be possible to identify emergent properties of the combined modalities, which are not present when used separately.

Chapter 8

Conclusion

Determining which interaction modalities to implement in an AR or VR system is a complex issue with many interconnected facets. Our work comparing three pure modalities in both speed and user experience contributes to the existing body of knowledge by illuminating aspects of these three pure modalities.

There are several conclusions we can draw from the work.

- The efficiency of the interaction and the users' experience with using it may not be as closely tied as we expected. We saw from the user study that even though the finger modality was significantly faster than the voice modality, using voice led to lower physical strain among the participants. Likewise, the participants' preferred modality was evenly split between finger and voice. This indicates a need to examine which parameters users value more highly, as well as more work in various combinations of modalities to discover superior permutations.
- Both the finger and hand modalities had a faster task completion speed than the voice modality. This may to some extent be a byproduct of the latency of the speech recognition in the prototype, but it is also likely that "direct" interactions such as pointing at the desired target are generally faster than more indirect methods, such as speech when completing non-abstract tasks.
- A short-duration, fully VR study like the one in this paper may hide some aspects of the participants' preferences. The shortness of the study may have only revealed the participants' initial impressions of the modalities. The participants may not have considered the implications of using the modalities for extended periods, or in more dynamic environments surrounded by other people.
- Finger-pointing is well suited for menu-based selection with the fastest completion speed and high user preference. Furthermore, voice control in combination with the head pointing worked well for menu-based selection, as it ranked just as high as finger pointing in user satisfaction despite low task completion speed. However, the somewhat simple tasks used in this study did not allow for much variance in how well each modality handled a specific type of task. A more detailed study would need to implement more complex tasks and examine the

efficiency and usability of each modality for different challenges. Additionally, we focused entirely on menu-based selection in which the pointer methods performed well. While voice also performed admirably, we did not utilize or test its strengths in the completion of more abstract tasks, which were absent from the study.

Acknowledgements

We want to thank our advisor Ken Pfeuffer for his help and guidance with the Bachelor's project. We also want to thank Mathias Lystbæk for his help given throughout the project. Lastly, we want to thank Pavel Manakhov for allowing us to use their asynchronous logging script which was used to record task completion times on the headset, as well as the friends and family who helped proofread the paper.

Bibliography

[1] Paul Milgram, Haruo Takemura, Akira Utsumi, and Fumio Kishino. Augmented reality: A class of displays on the reality-virtuality continuum. *Telemanipulator and Telepresence Technologies*, 2351, 01 1994.

[2] Jari Kaivo-oja, Theresa Lauraeus, and Mikkel Stein Knudsen. Picking the ict technology winners-longitudinal analysis of 21st century technologies based on the gartner hype cycle 2008-2017: trends, tendencies, and weak signals. *International Journal of Web Engineering and Technology*, 15(3):216–264, 2020.

[3] Peter Fernandez. Facebook, meta, the metaverse and libraries. *Library Hi Tech News*, 2022.

[4] Stylianos Mystakidis. Metaverse. *Encyclopedia*, 2(1):486–497, 2022.

[5] Apple Inc. Apple vision pro. `https://www.apple.com/apple-vision-pro/`. Accessed 07-06-2023.

[6] Dimensions. Dimensions. `https://app.dimensions.ai/discover/publication?search_mode=content&search_text=Virtual%20Reality&search_type=kws&search_field=full_search`. Accessed 24-05-2023.

[7] Jens Grubert, Tobias Langlotz, Stefanie Zollmann, and Holger Regenbrecht. Towards pervasive augmented reality: Context-awareness in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1706–1724, June 2017.

[8] Kristina Berntsen, Ricardo Colomo Palacios, and Eduardo Herranz. Virtual reality and its uses: A systematic literature review. In *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality*, TEEM '16, page 435–439, New York, NY, USA, 2016. Association for Computing Machinery.

[9] Richard A. Bolt. "put-that-there": Voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.*, 14(3):262–270, jul 1980.

[10] Adam Kendon. Some relationships between body motion and speech. *Studies in dyadic communication*, 7(177):90, 1972.

[11] Mitchell Baxter, Anna Bleakley, Justin Edwards, Leigh Clark, Benjamin R. Cowan, and Julie R. Williamson. "you, move there!": Investigating the impact of feedback on voice control in virtual environments. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[12] Becky Spittle, Maite Frutos-Pascual, Chris Creed, and Ian Williams. A review of interaction techniques for immersive environments. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2022.

[13] T.P. Caudell and D.W. Mizell. Augmented reality: an application of heads-up display technology to manual manufacturing processes. In *Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences*, volume ii, pages 659–669 vol.2, 1992.

[14] Julia Hertel, Sukran Karaosmanoglu, Susanne Schmidt, Julia Bräker, Martin Semmann, and Frank Steinicke. A taxonomy of interaction techniques for immersive augmented reality based on an iterative literature review. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 431–440, Oct 2021.

[15] Ivan Poupyrev, Tadao Ichikawa, Suzanne Weghorst, and Mark Billinghurst. Egocentric object manipulation in virtual environments: empirical evaluation of interaction techniques. In *Computer graphics forum*, volume 17, pages 41–52. Wiley Online Library, 1998.

[16] Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. Speech-augmented eye gaze interaction with small closely spaced targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, ETRA '06, page 67–72, New York, NY, USA, 2006. Association for Computing Machinery.

[17] Mathias N. Lystbæk, Peter Rosenberg, Ken Pfeuffer, Jens Emil Grønbæk, and Hans Gellersen. Gaze-hand alignment: Combining eye gaze and mid-air pointing for interacting with menus in augmented reality. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA), may 2022.

[18] Mathias N. Lystbæk, Ken Pfeuffer, Jens Emil Sloth Grønbæk, and Hans Gellersen. Exploring gaze for assisting freehand selection-based text entry in ar. *Proc. ACM Hum.-Comput. Interact.*, 6(ETRA), may 2022.

[19] Mikko Kytö, Barrett Ens, Thammathip Piumsomboon, Gun A. Lee, and Mark Billinghurst. Pinpointing: Precise head- and eye-based target selection for augmented reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.

[20] Aunnoy Mutasim, Anil Batmaz, and Wolfgang Stuerzlinger. Pinch, click, or dwell: Comparing different selection techniques for eye-gaze-based pointing in virtual reality. 05 2021.

[21] Mohamed Khamis, Florian Alt, and Andreas Bulling. Challenges and design space of gaze-enabled public displays. In *Proceedings of the 2016 ACM International*

Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, UbiComp '16, page 1736–1745, New York, NY, USA, 2016. Association for Computing Machinery.

[22] Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. Artention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers & Graphics*, 95:1–12, 2021.

[23] Linda E. Sibert and Robert J. K. Jacob. Evaluation of eye gaze interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, page 281–288, New York, NY, USA, 2000. Association for Computing Machinery.

[24] David Fono and Roel Vertegaal. Eyewindows: Evaluation of eye-controlled zooming windows for focus selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, page 151–160, New York, NY, USA, 2005. Association for Computing Machinery.

[25] Jonas Blattgerste, Patrick Renner, and Thies Pfeiffer. Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In *Proceedings of the Workshop on Communication by Gaze Interaction*, COGAIN '18, New York, NY, USA, 2018. Association for Computing Machinery.

[26] Yuan Yuan Qian and Robert J. Teather. The eyes don't have it: An empirical comparison of head-based and eye-based selection in virtual reality. In *Proceedings of the 5th Symposium on Spatial User Interaction*, SUI '17, page 91–98, New York, NY, USA, 2017. Association for Computing Machinery.

[27] Robin Piening, Ken Pfeuffer, Augusto Esteves, Tim Mittermeier, Sarah Prange, Philippe Schröder, and Florian Alt. Looking for info: Evaluation of gaze based information retrieval in augmented reality. In *Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30 – September 3, 2021, Proceedings, Part I*, page 544–565, Berlin, Heidelberg, 2021. Springer-Verlag.

[28] Feiyu Lu and Doug A Bowman. Evaluating the potential of glanceable ar interfaces for authentic everyday uses. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pages 768–777. IEEE, 2021.

[29] Feiyu Lu, Shakiba Davari, Lee Lisle, Yuan Li, and Doug A. Bowman. Glanceable ar: Evaluating information access methods for head-worn augmented reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 930–939, March 2020.

[30] Shakiba Davari, Feiyu Lu, and Doug A. Bowman. Validating the benefits of glanceable and context-aware augmented reality for everyday information access tasks. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 436–444, March 2022.

[31] R. Azuma and C. Furmanski. Evaluating label placement for augmented reality view management. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 66–75, Oct 2003.

[32] Knut Hartmann, Timo Götzelmann, Kamran Ali, and Thomas Strothotte. Metrics for functional and aesthetic label layouts. In Andreas Butz, Brian Fisher, Antonio Krüger, and Patrick Olivier, editors, *Smart Graphics*, pages 115–126, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[33] Jacob Boesen Madsen, Markus Tatzgern, Claus B. Madsen, Dieter Schmalstieg, and Denis Kalkofen. Temporal coherence strategies for augmented reality labeling. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1415–1423, April 2016.

[34] David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. Context-aware online adaptation of mixed reality interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, UIST '19, page 147–160, New York, NY, USA, 2019. Association for Computing Machinery.

[35] João Marcelo Evangelista Belo, Mathias N. Lystbæk, Anna Maria Feit, Ken Pfeuffer, Peter Kán, Antti Oulasvirta, and Kaj Grønbæk. Auit – the adaptive user interfaces toolkit for designing xr applications. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery.

[36] Chulmo Koo, Seunghun Shin, Ulrike Gretzel, William Hunter, and Namho Chung. Conceptualization of smart tourism destination competitiveness. *Asia Pacific Journal of Information Systems*, 26:561–576, 12 2016.

[37] Aurkene Alzua-Sorzabal, María Teresa Linaza, and Marina Abad. An experimental usability study for augmented reality technologies in the tourist sector. In Marianna Sigala, Luisa Mich, and Jamie Murphy, editors, *Information and Communication Technologies in Tourism 2007*, pages 231–242, Vienna, 2007. Springer Vienna.

[38] David Ingram. Trust-based filtering for augmented reality. In Paddy Nixon and Sotirios Terzis, editors, *Trust Management*, pages 108–122, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[39] Unity Technologies. Unity. `https://unity.com/`. Accessed 31-05-2023.

[40] Oculus Integration SDK downloads. `https://developer.oculus.com/downloads/package/unity-integration/m`. Accessed: 17-05-2023.

[41] 255 Pixel Studios. City package. `https://assetstore.unity.com/packages/3d/environments/urban/city-package-107224`, Dec 2020. Accessed: 19-05-2023.

[42] ChermandirKun. Horse statue. `https://assetstore.unity.com/packages/3d/environments/fantasy/horse-statue-52025`, Oct 2016. Accessed: 19-05-2023.

[43] ChamferBox Studio. Discobolus statue. https://assetstore.unity.com/packages/3d/props/discobolus-statue-107544, Jan 2018. Accessed: 19-05-2023.

[44] Wikipedia. Discobolus. http://en.wikipedia.org/w/index.php?title=Discobolus&oldid=1136682982, Jan 2023. Accessed 19-05-2023.

[45] Meta. wit.ai. https://wit.ai/. Accessed 07-06-2023.

[46] Martin Mitrevski. *Getting Started with Wit.ai*, pages 143–164. Apress, Berkeley, CA, 2018.

[47] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage publications Sage CA: Los Angeles, CA, 2006.

[48] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988.

[49] Lei Gao. Latin squares in experimental design. *Michigan State University*, 2005.

[50] James V. Bradley. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528, 1958.

[51] IBM. Ibm spss statistics. https://www.ibm.com/products/spss-statistics. Accessed 26-05-2023.

[52] Lærd Statistics. Sphericity. https://statistics.laerd.com/statistical-guides/sphericity-statistical-guide.php, 2018. Accessed 24-05-2023.

[53] Hervé Abdi et al. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3(01):2007, 2007.

[54] Michael R Sheldon, Michael J Fillyaw, and W Douglas Thompson. The use and interpretation of the friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International*, 1(4):221–228, 1996.

[55] Robert F Woolson. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3, 2007.

[56] Krystal Hu. Chatgpt sets record for fastest-growing user base. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, Feb 2023. Accessed 25-05-2023.

[57] Laura Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, 35:379–383, 2003.

Appendix A

Study results

A.1 TLX-questionnaire

Head pointer questionnaire:

https://forms.office.com/Pages/DesignPageV2.aspx?subpage=design&FormId=Nh39Ycv-yke319DfA3ChmK4g2EosbEVPnNAvYBUKu_lUOE83SEhBMUZGRk44QVRPS1kyTFlMTzE2Sy4u&Token=fd3b542235774e36b1b12e18f388f13a

Finger pointer questionnaire:

https://forms.office.com/Pages/DesignPageV2.aspx?subpage=design&FormId=Nh39Ycv-yke319DfA3ChmK4g2EosbEVPnNAvYBUKu_lUQUtTVjdZUzNSQTBKSTY1STREM0xQV1lERy4u&Token=db92582fbe834a97a3c853821895b78b

Voice command questionnaire:

https://forms.office.com/Pages/DesignPageV2.aspx?subpage=design&FormId=Nh39Ycv-yke319DfA3ChmK4g2EosbEVPnNAvYBUKu_lUMlU5Nzk3WTVVWDNWWk9YTk1VNThJUzZPVC4u&Token=2d4454f4baa045e6811fab6a6328fe19

Preferred modality question:

https://docs.google.com/document/d/10z_IOYhWXOcpHJREBL845gVFOHnrZoIbHkeyKyC6WYM/edit?usp=sharing

A.2 Task Completion Time

The task completion time for each user: https://drive.google.com/drive/folders/1mmQryp1vI3IfB5IIIeRI9zouxX8yb46p

Appendix B

The Technical Details

B.1 GitLab links

GitLab project: `https://gitlab.au.dk/ubi-onebacholor/dynamic-ui-for-eye-based-interaction-in-vr`

GitLab Data analysis: `https://gitlab.au.dk/ubi-onebacholor/bachlordatabehandling`

B.2 Wit.ai configuration

Wit.ai understanding: `https://wit.ai/apps/6242201739172058/understanding`